Data Mining

Part 2. Data Understanding and Preparation

2.2 Data Preparation

Spring 2010

Instructor: Dr. Masoud Yaghini

Outline

- Why Data Preparation?
- Major Tasks in Data Preparation
- References

Why Data Preparation?

Why Data Preparation?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”, martial status =“ ”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing inconsistencies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., inconsistency between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - It was not considered important at the time of entry
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

Why Is Data Preparation Important?

- No quality data, no quality mining results!
- Quality decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.

Major Tasks in Data Preparation

Major Tasks in Data Preparation

- Major Tasks in Data Preparation
  - Data cleaning
  - Data transformation

Major Tasks in Data Preparation

- Data cleaning
  - Fill in missing values
  - Identify or remove outliers
  - Resolve inconsistencies
  - Schema integration
  - Handling redundancy

Major Tasks in Data Preparation

- Data transformation
  - Normalization
  - Attribute construction (or feature construction)
  - Aggregation
  - Discretization
  - Generalization

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 2)

The end